Análisis de calidad de vídeo H.264 en streaming sobre HSUPA

Antonio Foncubierta Rodríguez[(1)], Ramón Cerquides Bueno[(1)]
antonio.foncubierta@gmail.com, cerquides@us.es
[(1)]Dpto. de Teoría de la Señal Universidad de Sevilla.

***Resumen*—En un panorama en el que cada vez se transmiten más contenidos multimedia a través de Internet, es necesario obtener de algún modo métricas que indiquen la calidad objetiva del material cuando el envío del mismo se hace bajo fuertes limitaciones de ancho de banda. El sistema de codifcación H.264-AVC es conocido por su excelente comportamiento en relación calidad - tasa de bits. A partir de envíos en streaming de diferentes vídeos codifcados con éste codec, estudiaremos cómo afectan las restricciones de régimen binario y la pérdida de paquetes a la calidad del vídeo recibido. Para ello se utilizarán distintas métricas (PSNR, SSIM, VQM), tratando así de modelar tanto el comportamiento objetivo del sistema como la calidad percibida por los usuarios, ev medidas de tipo MOS que puedan resultar costosas económica y temporalmitando la necesidad de realizar medidas de tipo MOS que puedan resultar costosas económica y temporalmente.**

I. Introducción

Durante cualquier procesado de imágenes digitales, éstas están sujetas a diferentes tipos de distorsión y fuentes de error, que hacen que la calidad de las mismas disminuya respecto al original. Los algoritmos con mayor ratio de compresión suelen ser con pérdidas, lo que significa que la imagen no puede recuperarse nunca íntegramente, variando la información que aporta. Cuando las imágenes son percibididas por humanos, la información no solo consiste en los píxeles representados, sino que la evaluación depende en gran medida del sistema visual humano y de otros factores como la propia experiencia del espectador.

Las técnicas de evaluación de calidad de imagen suelen dividirse en subjetivas y objetivas. Las primeras son normalmente medidas basadas en MOS (*Mean Opinion Score*) que consisten en promediar las opiniones de una muestra representativa. Las medidas objetivas suelen basarse en algoritmos que de forma automática evalúan las diferencias (aunque no en todos los casos) entre dos imágenes. Los costes derivados de realizar un experimento tipo MOS son altísimos en lo que a tiempo se refiere, puesto que es necesario mostrar las imágenes a cada uno de los expertos que van a evaluarla. No obstante, si la muestra está bien escogida, esta evaluación puede ser una medida bastante verosímil de la calidad de la imagen. La ventaja de los métodos objetivos es el ahorro en recursos que supone la evaluación de forma automática. Esto abre la puerta a diferentes aplicaciones, como monitorización dinámica de calidad de una secuencia de imágenes, optimización de parámetros de compresión para maximizar la calidad o simplemente la evaluación de calidades a posteriori. Debido a la forma en la que los humanos perciben las imágenes, la correlación entre los resultados aportados por pruebas subjetivas y objetivas no siempre coinciden, aunque en cualquier caso existe una cierta correlación entre ambos grupos.

Las técnicas más modernas de evaluación de calidad intentan modelar el sistema de percepción humano para, de forma matemática, poder predecir el resultado de una experiencia subjetiva. A lo largo de los últimos años, se han investigado y desarrollado diferentes técnicas en esta materia.

En el apartado II de la ponencia se describe el estado actual de las técnicas de transmisión y codificación que se emplean en el estudio, mientras que en el apartado III se discutirán las distintas métricas empleadas para evaluación de video. Por último, en los apartados IV y V se presentarán los resultados y se esbozarán las líneas futuras para continuar la investigación.

II. Escenario

El caso de estudio para el cual se requieren medidas de calidad de vídeo es el de un sistema de transmisión de vídeo de definición estándar comprimido usando H.264 [1] y enviado en tiempo real sobre una conexión de datos basada en la tecnología HSUPA[2]. Las caracteristicas propias de una conexión inalámbrica producen que sea necesario recurrir a técnicas de *error concealment* para disminuir la visibilidad de los errores en el flujo de datos y de técnicas avanzadas de comrpesión de vídeo para adaptar

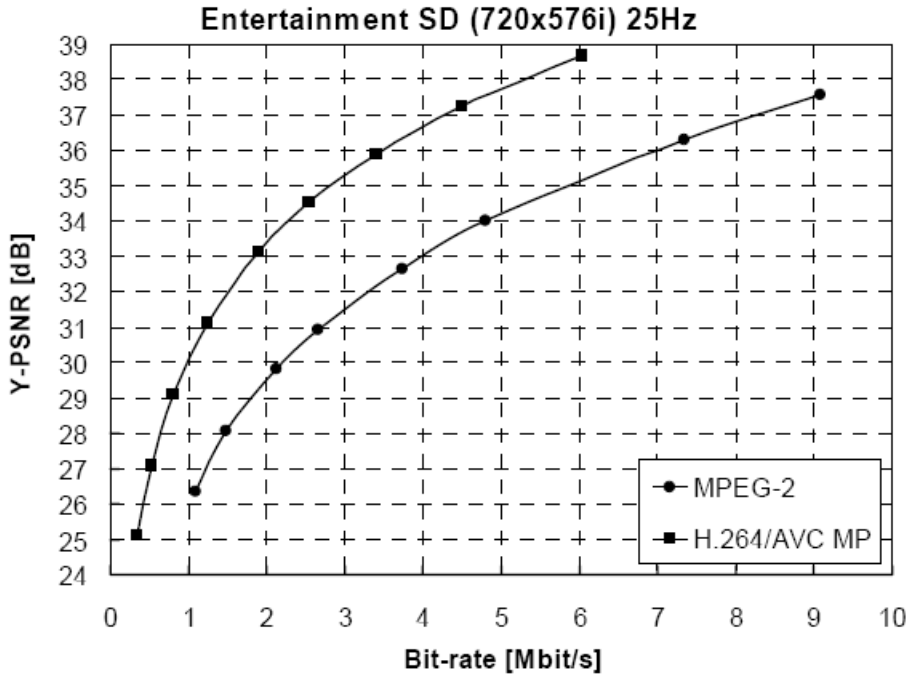

Figura 1. Comparación entre MPEG-2 y H.264

la tasa de transferencia a los límites soportados por la tecnología de acceso.

El sistema sujeto a estudio, consiste en la conexión de una cámara de vídeo digital profesional, con conexión FireWire, a un PC portátil donde la señal de vídeo es recodificada usando el perfil *Baseline* de H.264 a tasas entre 768 y 1024 Kbps, para su inclusión junto con un canal de audio monoaural codificado a 96 Kbps en un *Transport Stream* de MPEG. El flujo de vídeo se encapsula en paquetes RTP que viajan a través de la red de paquetes hasta el destino, utilizando para ello el punto de acceso a Internet proporcionado por Vodafone al terminal HSUPA. No obstante, para poder realizar medidas de calidad comparando diferentes versiones de un mismo vídeo, se utilizan archivos de video en lugar de capturar directamente desde una cámara digital.

En el destino, se monitoriza el flujo de datos recibidos, teniendo en cuenta los retrasos entre los paquetes, las variaciones del mismo y la cantidad de paquetes perdidos. Además, para la evaluación mediante técnicas objetivas de la calidad del vídeo, el flujo RTP puede ser salvado como un archivo de vídeo para su posterior visionado.

H.264

El sistema de compresión elegido se caracteriza por una excelente relación calidad-tasa de bits comparado con otros estándares actualmente en uso. Como puede apreciarse en la Figura 1, para una misma calidad de la señal en términos de PSNR, el estándar H.264 utiliza la mitad de ancho de banda que MPEG-2, que se usa por ejemplo en la codificación de DVD-Video y DVB-T (Televisión Digital Terrestre). En este último caso cada canal (*Program Stream*) consiste en una señal MPEG-2 codificada a 3Mbps, que utilizando H.264 sería poco más de 1 Mbps para la misma calidad.

H.264 aprovecha todos los tipos de redundancia en un flujo de vídeo (espacial, temporal y estadística) para reducir el régimen binario de la señal de vídeo perdiendo la mínima calidad posible. Su implantación comercial es cada vez más común, siendo el estándar de codificación elegido para vídeo de alta definición y para la televisión digital móvil. El estándar, desarrollado por el grupo de expertos MPEG en colaboración con la ITU, fue publicado en 2003, si bien se espera una mejora de su rendimiento al menos hasta 2013, ya que los codificadores comerciales actuales no explotan todo el potencial del sistema.

El perfil Baseline, usado en este estudio, nos proporciona las siguientes herramientas:

- I *Slices*.
- P *Slices*.
- CAVLC: *Context Adaptative Variable*

Length Coding.

- Agrupación de Slices y ordenación aleatoria de los mismos.
- *Slices* Redundantes.

Las principales diferencias de los métodos incorporados por H.264 respecto a MPEG-2 son por un lado la Intra predicción, esto es, que en lugar de enviarse una imagen comprimida en JPEG para los fotogramas tipo Intra, se estiman los macrobloques a partir de los macrobloques anteriores, permitiendo al codificador enviar sólo la información que permite corregir el error de esta estimación. Se logra asi una disminución notable de la tasa de bits consumida por dichos fotogramas, que eran los que menos compresión llevaban en MPEG2. Un ejemplo de la imagen que se logra por Intra Predicción antes de corregir el error puede verse en la figura 2. Por otro lado, se extiende la idea de predicción de fotogramas, y en lugar de codificar un fotograma a partir de un fotograma anterior o uno posterior, el codificador establece toda una lista de fotogramas de referencia anteriores y posteriores desde los cuales puede predecir cualquier otro fotograma. Por último se añade un filtro *deblocking* para evitar artefactos en la codificación.

HSUPA (High Speed Uplink Packet Access)

El estándar de telefonía de tercera generación UMTS, en su revisión de 2006, aporta mejoras sustanciales para la transmisión de tráfico de paquetes desde el terminal móvil a la red. Estas mejoras, junto a las introducidas en *UMTS Release 05 (HSDPA)*[3] vienen a proporcionar prestaciones similares a las de una conexión fija. El uso combinado de estas tecnologías, HSPA (High Speed Packet Access), es conocido comercialmente como Banda Ancha 3G, 3,75G o 3G+.

HSUPA utiliza mejoras en el canal ascendente similares a las aportadas por el estándar HSDPA(*UMTS Release 05*) en el descendente, como son:

- menores intervalos de transmisión que pasan de estar entre 80 y 10 ms a colocarse entre 10 y 2 ms
- mecanismos de retransmisión con redundancia incremental
- traslación de procesos de control de terminal al propio nodo B o estación base, con lo que la mejora de retrasos, latencia y velocidad de transmisión es sustancial
- procesos dinámicos de asignación de tasa de bits a los distintos terminales en una misma celda, que asignan una tasa en cada intervalo de 10 ms en función de las condiciones del canal (ruido), la potencia que puede transmitir el terminal y la cantidad de datos por transmitir

Categoría	**Máx. Velocidad**
Cat. 1	0,73 Mbps
Cat. 2	1,46 Mbps
Cat. 3	1,46 Mbps
Cat. 4	2,93 Mbps
Cat. 5	2,00 Mbps
Cat. 6	5,76 Mbps
Cat. 7	(3GPP Rel7) 11,5 Mbps

Cuadro I
CATEGORÍAS DE LOS TERMINALES HSUPA

HSUPA define distintas categorías para los terminales, según las prestaciones que pueden proporcionar. En el cuadro I se pueden observar las distintas categorías y la tasa máxima de subida soportada. Se ha incluido la actualización de las especificaciones de HSUPA, para la séptima revisión de UMTS, publicada recientemente y que incrementa el límite máximo de subida hasta los 11.5 Mbps. Actualmente la red tiene instalada equipos que soportan Categoría 2, mientras que los terminales que se comercializan, como por ejemplo el Huawei E272 soportan Categoría 5.

La especificación del 3GPP considera determinadas medidas de Calidad de Servicio para la tecnología HSUPA. Estas medidas irían desde la clasificación de clientes (oro, platino, etc.) hasta la especificación de tráfico no sujeto al planificador (*scheduler*). Este último sería previsiblemente tráfico de baja intensidad. Sin embargo, ninguna de estas medidas se ha puesto aún en práctica por parte de las operadoras, en parte debido a que los fabricantes de equipo no las incorporan todavía a sus dispositivos. El año 2009 se perfila como el momento en el que se tomarán medidas para la implantación de QoS.

III. MÉTRICAS DE CALIDAD

A lo largo del estudio se han realizado transmisiones de diferentes vídeos y se ha evaluado de diversas formas la calidad de los mismos de forma comparativa, es decir, para cada fotograma, se comparan la imagen de la secuencia original con la de la secuencia recibida. Las medidas que se han utilizado han sido la Relación Señal a Ruido de Pico (PSNR) como medida no perceptual; y dos medidas que procuran predecir la percepción de la calidad basándose en diferentes técnicas: SSIM y VQM.

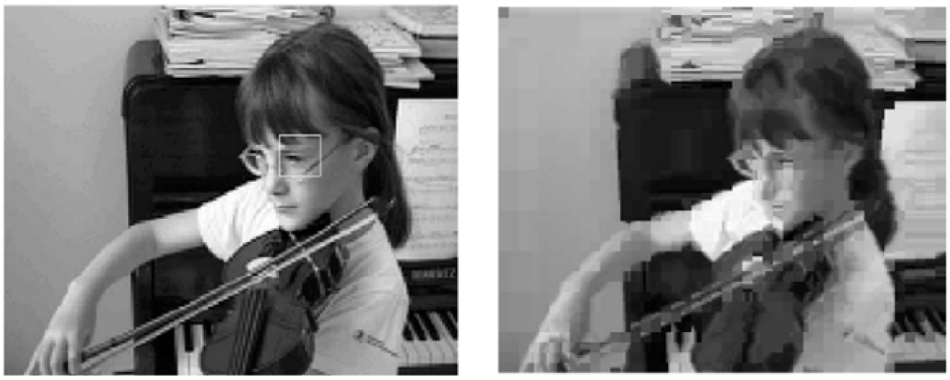

Figura 2. Resultado de la estimación por Intrapredicción

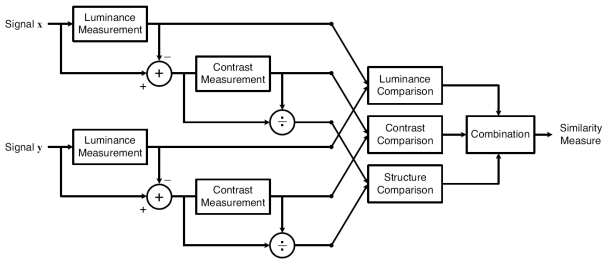

Figura 3. Diagrama de un sistema de medida basado en la similitud estructural.

Figura 4. Imagen original y cinco copias con el mismo Error Cuadrático Medio (MSE=210).

PSNR

Esta medida es muy usada en la práctica para muchas otras aplicaciones en ingenieria, sin embargo, al ser una técnica no perceptual sino totalmente objetiva los resultados que produce no se ajustan completamente a los que cabría esperar de un estudio tipo MOS. En sentido amplio, coincide con el Error Cuadratico Medio (MSE) pero, la propiedad logarítmica (ver Ecuación 1) de la PSNR la hace más conveniente. La métrica puede aplicarse a cada uno de las componentes de una imagen en color, pero dado que el sistema visual humano es mucho más sensible las variaciones de luminosidad que a las de color, se optó por aplicarla exclusivamente a la componente de luminancia (Y) del espacio de color YUV.

$$d(X,Y) = 10\log_{10}\frac{255^2 nm}{\sum_{i=1,j=1}^{n,m}(X_{i,j}-Y_{i,j})^2} \quad (1)$$

SSIM

La métrica SSIM[4] consiste en procurar independizar factores como la luminosidad y el contraste del resto de la imagen para comparar

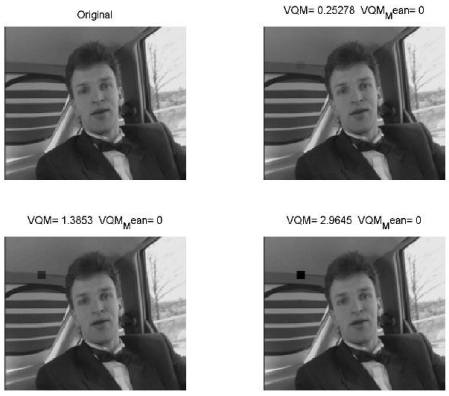

Figura 5. Imágenes con distinto VQM y media cero.

de forma estructural los objetos que aparecen en la misma (Figura 3). El desarrollo de éste método surge por la necesidad de tener alguna medida más allá del Error Cuadratico Medio o la PSNR, ya que el comportamiento de estas últimas métricas no proporcionan suficiente información sobre los cambios que hay en la imagen, como puede observarse en la Figura 4.

VQM

La métrica VQM descrita por Xiao [5] se basa en el uso de la Transformada Discreta Coseno para aprovechar las propiedades espacio-temporales del sistema de percepción humano. A su vez, Xiao se basa en la propuesta de Watson[6], añadiendo algunas mejoras como:

- Aplicar dos matrices SCSF (Función de Sensibilidad a Contraste Espacial): una para imágenes fijas y otra para imágenes dinámicas, basadas a su vez en las matrices de cuantización MPEG. La diferencia entre ambas es una operación potencia proporcional a la tasa de fotogramas.
- Se estima el valor de la métrica a partir del máximo valor de distorsión en cada fotograma, ponderado con la media de distorsión. Según el autor, esto refleja el hecho de que una distorsión grande localizada enmascara la percepción de pequeñas distorsiones en el resto del fotograma (Figura 5).

IV. Resultados Obtenidos

Los resultados obtenidos se basan en la aplicación de las técnicas mencionadas a tres vídeos diferentes, en varias ocasiones, y con distintas tasas de codificación. Estas medidas muestran una dependencia de la calidad mucho más fuerte con las caracteristicas del video que con la tasa de codificación. Habida cuenta de que la

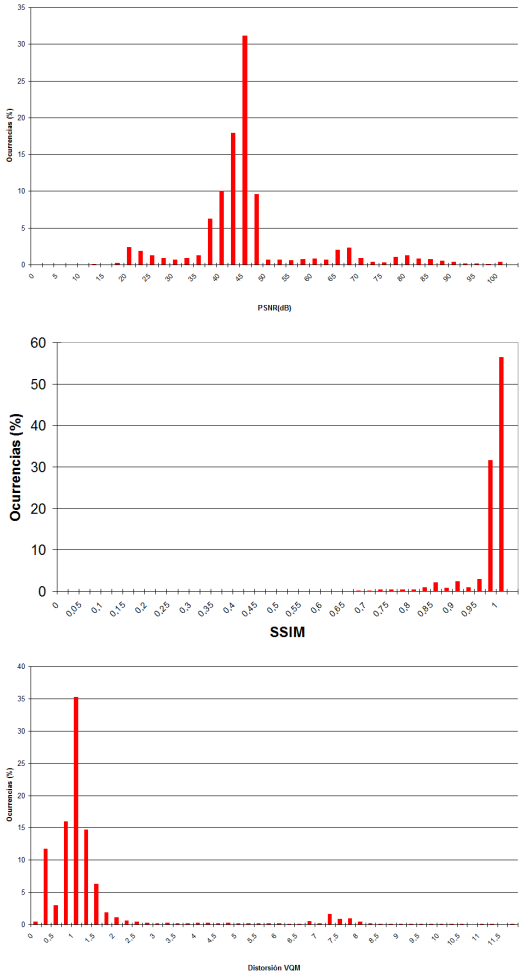

Figura 6. Vídeo Sencillo - Datos Empíricos.

tecnología permite un margen relativamente pequeño de mejora de la tasa de codificación (nos movemos entre 768 y 1024 Kbps) esta puede ser la causa de que la dependencia sea mucho mayor con las características del vídeo.

Vídeo Sencillo

Se analizan a continuación las pruebas realizadas con material sencillo, con pocos cambios de plano y con movimientos lentos y suaves. Las características del material permiten al codificador una eficiencia muy alta a la hora de codificar en tiempo real. En la Figura 6 se muestra un análisis gráfico de los datos obtenidos. Como puede observarse, la mayor parte de los fotogramas tienen un parecido bastante alto con el original, ya que la media de la PSNR se sitúa en los 44 dB. Este parecido se aprecia de una forma mucho más notable en la medida de la Similitud estructural, cuyo valor medio es de 0,96 tantos por uno.

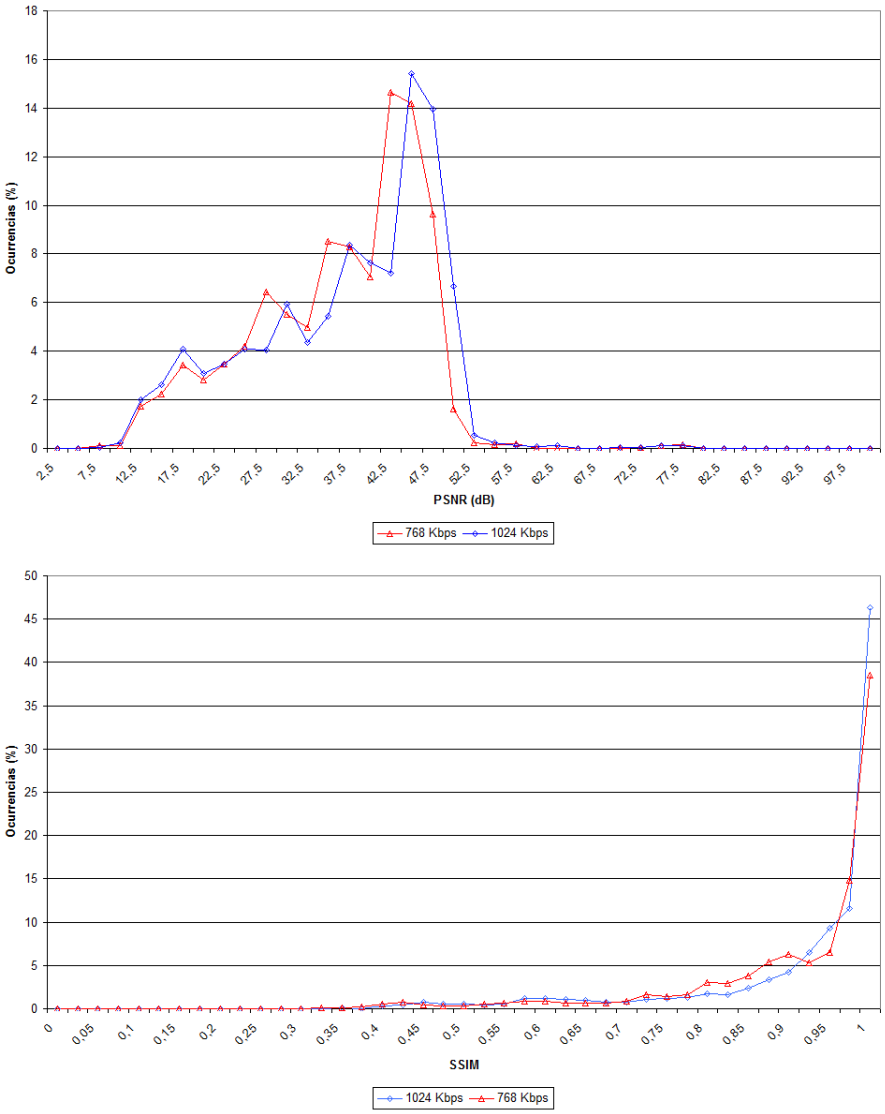

Figura 7. Vídeo Complejo, PSNR y SSIM - Datos Empíricos.

Vídeo Complejo

Mientras que con un vídeo sencillo apenas se encuentran problemas, y las pérdidas de paquetes si son leves permiten que las técnicas de ocultación de errores previstas por H.264 minimicen la visibilidad de los errores, al tratar con un vídeo más complejo, en el que las velocidades de los objetos son superiores, se comprueba un empeoramiento de la calidad (Figuras 7 y 8). En este caso, analizamos los resultados para dos calidades distintas de video un vídeo codificado a 768 Kbps y otro a 1024 Kbps. Como puede observarse la mejoría es pequeña al incrementar la tasa de codificación y las diferencias con respecto a un vídeo sencillo consisten en un ensanchamiento de la función densidad de probabilidad.

V. Líneas Futuras

La mejora de la capacidad computacional de los ordenadores permitirán que los codificadores basados en H.264 sigan aumentando su rendimiento durante los próximos años. El grupo de expertos encargado de realizar el estándar concibió un tiempo de vida de 10 años para H.264, alcanzando su máximo rendimiento en torno a 2013. Los resultados obtenidos en un estudio similar al presente proporcionarían por tanto datos mejores en cuanto a la calidad del video, por lo que la tecnología propuesta se perfila como una opción real para la realización de vídeo en exteriores.

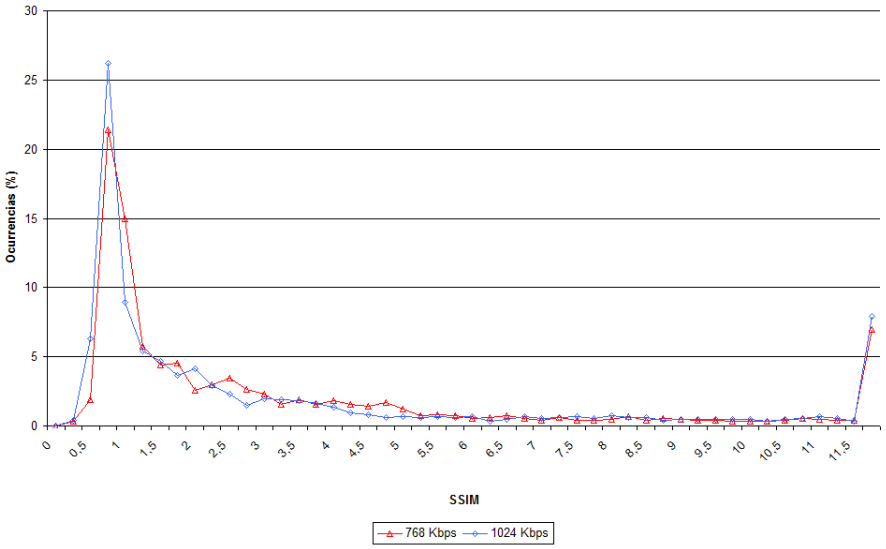

Figura 8. Vídeo Complejo, VQM - Datos Empíricos.

AGRADECIMIENTOS

Este proyecto ha sido financiado por la Consejería de Innovación y Desarrollo de la Junta de Andalucía, en el marco del Proyecto Minerva

REFERENCIAS

[1] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. Cited By (since 1996): 820.

[2] Universal mobile telecommunications system technical specifications and technical reports for a utran-based 3gpp system (3gpp ts 21.101 version 6.8.0 release 6). en línea: [ftp://ftp.3gpp.org/specs/2008-03/Rel-6/] Mayo 2008.

[3] Universal mobile telecommunications system technical specifications and technical reports for a utran-based 3gpp system (3gpp ts 21.101 version 5.13.0 release 5). en línea: [ftp://ftp.3gpp.org/specs/2008-03/Rel-5/] Mayo 2008.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. Cited By (since 1996): 333.

[5] Feng Xiao. Dct-based video quality evaluation.

[6] A.B. Watson. Toward a perceptual video quality metric. In *Human Vision, Visual Processing and Digital Display*, volume 3, pages 139–147, 1998.